M U N I

Annex No. 10 to the MU Directive on Habilitation Procedures and Professor Appointment Procedures

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University	
Applicant	Vít Nováček
Habilitation thesis	A Journey in Biomedical Discovery Informatics: From Ontology Learning to Knowledge Graph Embeddings
Reviewer	Prof. Stefan Schulz
Reviewer's home unit, institution	Medical University of Graz, Austria

The habilitation thesis presented by Vít Nováček encompasses about 14 years of academic work centred on topics on computational semantics using symbolic, probabilistic and neural approaches, with a focus on life science. The habilitation is cumulative and includes nine original publications, preceded by an introduction in which the term "Biomedical Discovery Informatics" is presented as the motto of his research activities.

The author divided his contributions into two blocks that characterise two phases of his research work, divided by a period in which he published little, according to his CV:

- The block "Ontology Learning" includes three journal articles with Vít Nováček as first author. However, two of these three papers were written prior to his PhD (2008 and 2010).
- The block "Knowledge Graph Embeddings" includes four journal articles and two proceedings article. Only in one conference article, Vít Nováček appears as First Author. His is last (senior) author in another journal article as well as in a proceeding article. In the remaining three articles of this block, he only appears as intermediate co-author, and his specific contribution is not specified by the publishers.

All these papers were peer-reviewed and appeared in proceedings of renowned conferences and journals with good impact factors for the domain. In each of these works Vít Nováček

demonstrates deep domain knowledge, a good command of research methodology and good publication skills.

According to the publication date and author order, I have concentrated my assessment mainly on his papers published after his PhD and those with him as first or last author, i.e. the publications on SKIMMR, BioKG, Kinase-substrate networks, and Biological applications of knowledge graphs. The consideration of these four papers only would be sufficient for the conclusion of my review.

In ***SKIMMR: Facilitating knowledge discovery in life sciences by machine-aided skim reading***, Vít Nováček exposes a system of semantic navigation of scientific article content by using semantic proximity (similarity and co-occurrence) link for contextualised content presentation. The methods are based on conventional distributional semantics, which is perhaps the reason that this work, published 2014 was not cited very often, in the view of the shift to neural approaches after 2015. Unfortunately, the web links provided in this paper that point to a demonstrator application and to the data are no longer active, which prevents the dataset being re-used, which would be interesting given the current state of the art of deep learning.

In ***BioKG: A knowledge graph for relational learning on biological data***, Vít Nováček is the senior author of a team that proposes a new biological knowledge graph pattern, which is then filled from established databases like Reactome, KEGG, Uniprot etc., using a script they provide to the community. Their approach is distinct from existing ones in which the interesting information (I would say at A-Box level, namely the data from the annotations) is too much dominated by ontological (T-Box) information. The authors also provide benchmarks, but do not explain in detail how they were created. I see this work on BioKG as a first step of several iterations, in which some more detail could be added to the representational template, (Fig. 1). Which, e.g., does not (yet) include entity types like cell component, molecular function (activity) and biological process from GO, as well as tissue, organism and phenotype. Also the conceptual difference between links, properties and metadata would require some more clarification, e.g. why drug side effects are under "property", drug protein interactions under „link", and drugs – ATC code associations are not under "metadata".

The paper ***Accurate prediction of kinase-substrate networks using knowledge graphs*** is undoubtedly the most significant work written by Vít Nováček as a first author. It describes a carefully designed experimental setting, is very well written and illustrated. I consider it an absolutely exemplary approach for link prediction in biology. The paper describes the design of LinkPhinder, a tool that predicts kinase activity, i.e. the attachment of a phosphate group

to a molecule, a very universal mechanism in biochemistry and therefore highly relevant for finding drug targets. It describes how existing phosphorylation data are transformed into a knowledge graph, out of which link prediction models were trained. The experimental validation against two datasets shows how the chosen approach outperforms the state-of-the art, sequence based approaches. Additional support for the link prediction model is also given by wet lab experimental confirmation of the newly hypothesised kinase – substrate associations. Software and data were made available to the public.

Among all works presented the paper ***Biological applications of knowledge graph embedding models*** appeared in the journal with the highest impact factor. As a senior author he summarised the state of the art on knowledge graph embedding (KGE) models in life science in form of a review, but also includes experimental sections that compare KGE models with each other and demonstrate the two main applications of KGE, namely prediction and clustering. The work is completed by the analysis of runtime behaviour of KGE approaches, as well as their strengths and weaknesses. It is a very well written, easy to read paper of educational value.

Reviewer's questions for the habilitation thesis defence:

- If you implemented SKIMMR again, which methodological choices would be different now, compared to the work done ten years ago
- When injecting axioms from ontologies into knowledge, which approach would you follow for equivalence axioms regarding class definitions, such as in OWL $A \equiv B \sqcap \exists r.C$ or more complex ones?
- For a Knowledge graph schema such as Fig. 1 in the last paper (Biological applications of KGE models), would you see some benefit to align it with foundational ontologies such as developed by the Applied Ontology community, which have also influenced the OBO foundry ontologies?
- Or would you say that formal ontologies as well as logics languages like those based on description logics would be mostly irrelevant for knowledge graphs, also, e.g. the division of knowledge graph nodes into A-Box and T-Box entities?
- After authoring your last paper, what has changed in the field of knowledge graph embeddings, particularly regarding neural KGE Models (cf. section 3 in your paper Regularizing Knowledge Graph Embeddings)?

Conclusion

The habilitation thesis entitled "*A Journey in Biomedical Discovery Informatics: From Ontology Learning to Knowledge Graph Embeddings*" by Vít Nováček **fulfils** the requirements expected of a habilitation thesis in the field of Informatics.

Date: April 3rd, 2022

Signature: